BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37306
doi: 10.3897/biss.3.37306

OPEN ACCESS

Conference Abstract

Evaluating Methods for Transcribing Specimen Labels

Sarah Phillips[‡], Mathias Dillen[§], Laura Green[‡], Quentin Groom[§], Marie-Helene Weech[‡]

‡ Royal Botanic Gardens Kew, Richmond, United Kingdom
§ Meise Botanic Garden, Meise, Belgium

Corresponding author: Sarah Phillips (sarah.phillips@kew.org)

Received: 14 Jun 2019 | Published: 21 Jun 2019

Citation: Phillips S, Dillen M, Green L, Groom Q, Weech M (2019) Evaluating Methods for Transcribing Specimen Labels. Biodiversity Information Science and Standards 3: e37306. https://doi.org/10.3897/biss.3.37306

Abstract

Distributed Systems of Scientific Collections (DiSSCo) a pan-European Research Infrastructure will facilitate the production of tens of millions of digital images of natural history specimens each year. The labels of these specimens contain valuable information for research, but their transcription can be difficult and time-consuming, with often hard to read handwritten labels. Whilst accurate label transcription is only one step along the way to create a specimen record fit for different research uses, it is an extremely important one. It would be very time-consuming to have to return to recheck label information for even a very small proportion of specimens. Once a specimen label is transcribed correctly, it becomes much easier to enhance the record with additional information from other sources, e.g. from literature or collector itineraries. It also becomes feasible to determine the point of collection from the textual information on the label by a process known as georeferencing, or even to find inaccuracies within the label itself.

Under the auspices of the project Innovation and Consolidation for Large Scale Digitisation of Natural Heritage (ICEDIG), we compared different manual approaches to transcription of collection labels. Using herbarium specimens as an example, the quality of transcribed data by:

1. in-house trained institute staff,
2. outsourcing to a commercial company or

3. transcription by the general public through online crowdsourcing platforms was compared through two transcription pilots.

The first pilot consisted of 200 S*olanum* specimen images from the Royal Botanic Gardens Kew in the UK and 200 from Meise Botanic Garden in Belgium. This particular genus was chosen as both institutes had specimens from which the label data had already been transcribed through the digitisation company Picturae, completed by Alembo. The Kew specimens had also been transcribed in-house by staff employed as digitisation officers or curators and by an independant researcher. The images from both institutes were uploaded to two crowdsourcing platforms: DigiVol and DoeDat. In a second pilot, multiple European institutions holding botanical collections were approached to provide a sample of 200 digitally imaged herbarium sheet specimens to upload to multiple crowdsourcing platforms. Specimens from 7 institutions were uploaded for transcription to 5 different crowdsourcing platforms: DigiVol, DoeDat, Die Herbonauten, Les Herbonautes and Notes from Nature.

For both pilots, key transcription data were assessed and common errors in label transcription identified. Reasons for these errors will be discussed along with possible mechanisms to improve the accuracy of the transcriptions. The need for standards for transcription is identified and recommendations made.

Keywords

data capture, data quality, Citizen Science

Presenting author

Sarah Phillips

Presented at

Biodiversity_Next 2019

Funding program

ICEDIG: Innovation and Consolidation for large-scale Digitisation of natural heritage

Grant title

ICEDIG: Innovation and Consolidation for large-scale Digitisation of natural heritage